Estimating ungauged streamflow using principal components

G. A. FULLER
Faculty of Engineering, University of Regina, Regina, Sask., Canada S4S 0A2

Received August 10, 1978
Accepted May 29, 1979

A principal components regression model for estimating ungauged streamflow data has been tested. This model overcomes the computational problem that may develop in a conventional least squares model when the determinant of the correlation matrix approaches zero. The parameters required by the model are estimates of the means and standard deviations of flows at the ungauged site and estimates of correlation coefficients between flows at the ungauged site and nearby hydrometric sites. These parameters were obtained from least squares models using physiographic characteristics as predictor variables. The model was tested by estimating streamflow records assumed to be nonexistent. These tests show that, while eliminating the possibility of the computational problem, the principal components model estimates are similar to those from a conventional regression model. Consequently, the principal components model is more reliable for highly correlated predictor variables. An example is given to illustrate the use of the principal components model.

On a mis à l'épreuve un modèle de régression des composantes principales (orthogonales) pour l'estimation des débits non jaugés. Le modèle surmonte les difficultés de calcul qui peuvent se présenter dans un modèle aux moindres carrés plus conventionnel lorsque le déterminant de la matrice de corrélation approche zéro. Les paramètres exigés par le modèle consistent dans les valeurs estimées des moyennes et des écarts types des débits au site non jaugé, et dans les valeurs estimées des coefficients de corrélation entre les débits au site non jaugé et aux stations hydrométriques voisines. Ces paramètres ont été établis au moyen d'un modèle aux moindres carrés utilisant les paramètres physiographiques comme variables dépendantes. On a vérifié le modèle en estimant des débits correspondant à des enregistrements supposés inexistants. De tels essais ont révélé que, tout en éliminant les risques de problèmes de calcul, les valeurs établies au moyen du modèle aux composantes principales sont semblables à celles qu'on obtient d'un modèle conventionnel de régression. On conclut que le modèle aux composantes principales est plus fiable pour des variables hautement dépendantes et son utilisation est illustrée par un exemple.
[Traduit par le revue]

Can. J. Civ. Eng., 6, 468–472 (1979)

Introduction

The least squares regression model, shown in [1], can be used to estimate streamflow data at sites for which no data exist using records from n nearby recording sites as the independent or predictor variables.

$$[1] \quad \boldsymbol{Y} = \alpha + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_i \boldsymbol{X}_i + \ldots + \beta_n \boldsymbol{X}_n + \boldsymbol{Z} s_Y (1 - r^2)^{1/2}$$

where $\boldsymbol{Y}$ is the vector of the estimates of ungauged flows; α and β_i are the least squares constant and coefficients respectively; $\boldsymbol{X}_i$ is a vector containing the streamflow records from site i; $\boldsymbol{Z}$ is a vector of normal standard deviates; s is the standard deviation; and r is the multiple correlation coefficient.

The least squares regression coefficients required in [1] can be calculated from estimates of the correlation coefficients between the flows at the ungauged site and nearby hydrometric sites. Physiographic characteristics can be used to estimate these correlation coefficients, as well as the means and standard deviations of the flows at the ungauged site (Fuller 1978). The relationship given in [2] can be used to determine the least squares coefficients required in [1].

$$[2] \quad \beta_i = (s_Y/s_{X_i})\,(m_{YX_i}/D_X)\,(-1)^{i+1}$$

where s_Y is the standard deviation for $\boldsymbol{Y}$, s_{X_i} is the standard deviation of the ith streamflow record; m_{YX_i} is the minor of the correlation matrix between Y and X's obtained by deleting the row associated with Y and the column associated with the ith streamflow record; and D_X is the determinant of the correlation matrix for the X's.

When using a least squares model such as the one shown in [1], problems can occur in computing the least squares coefficients if high correlations exist between the data from adjacent sites used as predictor variables (Young *et al.* 1970). As a result of the high intercorrelations the determinant of the predictor variable correlation matrix, which is used in computing the least squares coefficients, may approach zero. When this happens, it is difficult to obtain reliable least squares coefficients. To overcome this problem, the rank of the matrix of pre-

0315-1468/79/030468-05$01.00/0

dictor variables must be reduced so that the determinant does not approach zero. This report deals with a method of overcoming the problem of highly correlated data: the use of principal components analysis.

Principal Components Analysis

Principal components are an orthogonal transformation of a set of variables, in this case the streamflow data from the predictor stations. A set of n predictor variables is transformed to a set of n principal components as shown in [3].

$$[3] \quad \mathbf{E} = \mathbf{XA}$$

where **E** is a matrix containing vectors of principal components in columns; **X** is a matrix containing the vectors of the streamflow records in columns; and **A** is a matrix containing the eigenvectors of the correlation matrix of X in columns. The value of an element, e_{ji}, of the matrix **E** can be written as:

$$[4] \quad e_{ji} = \sum_{k=1}^{n} a_{ki} x_{jk}$$

The eigenvectors used in obtaining the principal components are chosen so that the first principal component explains the maximum possible amount of the total variance contained in the original predictor variables. Similarly, the second principal component explains the maximum amount of the remaining variance and so on. This means that the variance explained by the nth principal component is a minimum. Consequently, by discarding the nth principal component the rank is reduced and the remaining principal components explain the largest possible proportion of total variance of the original variables that can be explained by a reduced rank model. Thus, the computational problem is overcome by discarding one or more of the principal components and carrying out a least squares analysis on the remaining principal components. Since most computing centres have subroutines for computing principal components, this technique can easily be used to estimate streamflow data as shown by the following example.

Results

In order to exemplify the method of rank reduction using principal components, the site 02GA018 in southern Ontario is assumed to be ungauged and nearby stations 02GD012, 02HB001, 02HB002, 02GA003, and 02GA015 are used as predictors to estimate the flows for the period 1954–1968. The locations of these hydrometric sites are shown in Fig. 1. The monthly means, the monthly standard deviations, and the correlations with the predictor stations were estimated for site 02GA018 (Fuller 1978). The average errors in estimating the means, standard deviations, and correlation coefficients were 15, 20, and 5% respectively.

The first step in estimating the ungauged flow values at site 02GA018 is to standardize the predictor variable records in **X** giving them a mean of zero and a variance of 1.0. Then the correlation matrix **C** is determined for matrix **X**.

$$\mathbf{C} = \begin{bmatrix} 1.00 & 0.63 & 0.71 & 0.81 & 0.78 \\ 0.63 & 1.00 & 0.84 & 0.78 & 0.83 \\ 0.71 & 0.84 & 1.00 & 0.82 & 0.87 \\ 0.81 & 0.78 & 0.82 & 1.00 & 0.92 \\ 0.78 & 0.83 & 0.87 & 0.92 & 1.00 \end{bmatrix}$$

The roots of the correlation matrix (called eigenvalues) are obtained using a standard computing algorithm and then arranged in order of size. λ is the notation for the vector of eigenvalues. In this example:

$$[5] \quad \lambda = [4.20, 0.40, 0.17, 0.15, 0.07]$$

Associated with each eigenvalue is a vector called an eigenvector. The eigenvectors are computed along with the eigenvalues and stored in the columns of the matrix **A**. These columns of eigenvectors are arranged in the same order as their corresponding eigenvalues.

$$\mathbf{A} = \begin{bmatrix} 0.42 & -0.74 & -0.51 & 0.11 & 0.06 \\ 0.43 & 0.57 & -0.35 & 0.60 & -0.05 \\ 0.45 & 0.30 & -0.26 & -0.78 & -0.18 \\ 0.46 & -0.18 & 0.60 & 0.14 & -0.61 \\ 0.47 & 0.02 & 0.43 & -0.04 & 0.77 \end{bmatrix}$$

The first principal component vector is defined by [3] as:

$$[6] \quad E_1 = a_{11}X_1 + a_{21}X_2 + a_{31}X_3 + a_{41}X_4 + a_{51}X_5$$

Substituting the values from vector A gives:

$$[7] \quad E_1 = 0.42X_1 + 0.43X_2 + 0.45X_3 + 0.46X_4 + 0.47X_5$$

By definition the values in vector E_1 have a mean of 0 and a variance of λ_1 that equals 4.20. Since the total variance in the matrix **C** is equal to the sum of the diagonal elements of **C**, that is 5.0, the first principal component accounts for $(4.2/5.0) \times 100$ or 84% of the total variance in the matrix **C**.

Similarly, the second principal component vector

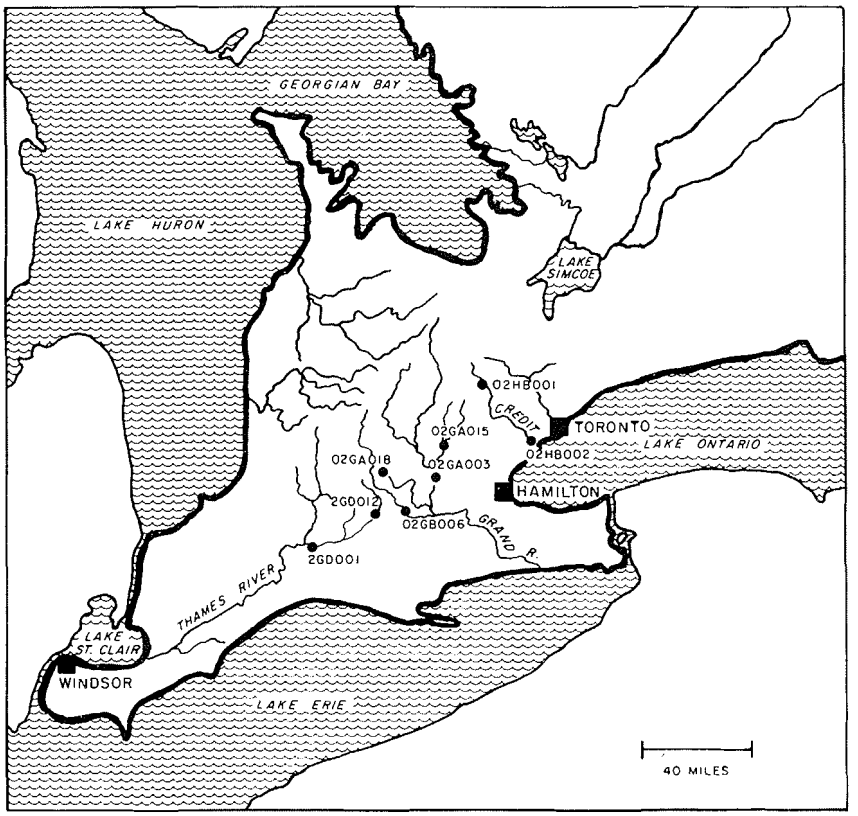

FIG. 1. Streamflow gauging site locations. Note: 1 mi = 1.6 km.

is:

$$[8] \quad E_2 = -0.74X_1 + 0.57X_2 + 0.30X_3 - 0.18X_4 + 0.02X_5$$

E_2 has a mean of 0 and a variance of 0.40. Thus, the first two principal components account for $[(4.2 + 0.4)/5.0] \times 100 = 92\%$ of the variance in the correlation matrix.

Continuing in a similar manner it is found that the first three principal components account for 95.4% of the variance. It was found in this study by estimating records assumed to be nonexistent that the best results from a reduced rank model were obtained when approximately 95% of the variance was explained. Therefore, since the first three principal components explain just over 95% of the variance, a regression analysis is performed on these first three principal components; therefore:

$$[9] \quad Y = a + b_1E_1 + b_2E_2 + b_3E_3 = a + b_1\mathbf{X}A_1 + b_2\mathbf{X}A_2 + b_3\mathbf{X}A_3$$

where A_i is the ith eigenvector from matrix $\mathbf{A}$. The coefficient a equals zero since the means of the variables equal zero. Since the principal components are independent, their correlation matrix is an identity or unit matrix. Since the principal components are uncorrelated and since the estimated values are assumed to have a standard deviation of 1.0 because the predictor variables were standardized, [2] becomes:

$$[10] \quad b_i = (s_Y/s_{E_i})\,(m_{YE_i}/D_E)\,(-1)^{i+1} = r_{YE_i}/\lambda_i^{1/2}$$

From the definition of a correlation coefficient and from [4] we get:

$$[11] \quad r_{YE_i} = \frac{\sum_{j=1}^{p} y_j e_{ji}}{p\lambda_i^{1/2}} = \frac{\sum_{j=1}^{p} y_j \sum_{k=1}^{n} a_{ki}x_{jk}}{p\lambda_i^{1/2}}$$

where p = the number of data in streamflow records. Equation [11] can be written as:

$$[12] \quad r_{YE_i} = \frac{\sum_{k=1}^{n} a_{ki} \sum_{j=1}^{p} y_j x_{jk}}{p\lambda_i^{1/2}} = \frac{\sum_{k=1}^{n} a_{ki} r_{YX_k}}{\lambda_i^{1/2}}$$

Substituting [12] into [10] gives the simplified equation used to determine the least squares coefficients on the principal components:

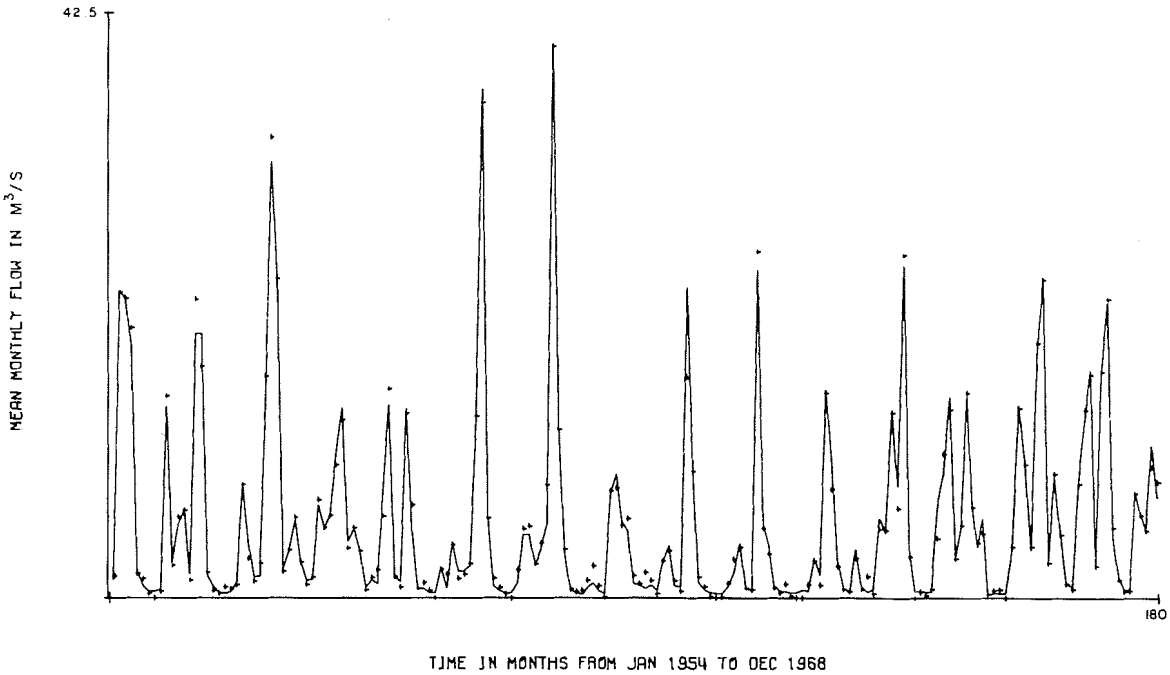

FIG. 2. Hydrograph and flow estimates for site 02GA018.

$$[13] \quad b_i = \frac{\sum_{k=1}^{n} a_{ki} r_{YX_k}}{\lambda_i}$$

After the least squares coefficients have been determined, estimates of the ungauged flows can be obtained from [9]. However, the estimates obtained from this equation will not include the random, unexplained component that naturally exists. For this reason a normal standard deviate, similar to that shown in [1], was added to the estimates obtained from [9] so as to increase the standard deviation of these estimates to 1.0. These standardized estimates are then "unstandardized" by giving the data the means and standard deviations that were estimated for the flow at site 02GA018. Depending upon the degree of model complexity desired, the estimated data can be "unstandardized" using estimated means and standard deviations for specific periods of the year such as months, which will result in the data from each month having a mean and standard deviation equal to the values estimated for that month. To give an indication of the reliability of the model for estimating monthly streamflow values, Fig. 2 shows the recorded hydrograph and estimated values for site 02GA018 in southern Ontario.

To compare the principal components model estimates with full rank estimates and actual streamflow records, pertinent statistics of the estimates and records are given in Tables 1, 2, and 3 for three streamflow records assumed to be nonexistent. These statistics are for the period 1954–1968 inclusive. It is evident from these tables that the differences between the full rank and reduced rank estimates are very small. Also, there was little difference between the computation times required by the two models. Consequently, the use of the reduced rank model is justified on the basis that the possibility of the occurrence of a computational problem is eliminated.

Conclusions

The accuracy of the estimates from the proposed

TABLE 1. Comparison between statistics of full rank and reduced rank estimates of flows* at site 02GB006

	Reduced rank estimates	Full rank estimates	Streamflow record
Mean	2.01	2.01	1.70
Standard deviation	2.46	2.46	1.98
Correlation with stn. 1	0.93	0.93	0.97
Correlation with stn. 2	0.73	0.73	0.75
Correlation with stn. 3	0.83	0.84	0.83
Correlation with stn. 4	0.80	0.80	0.83
Correlation with stn. 5	0.81	0.81	0.85
Serial correlation	0.38	0.38	0.41
Coefficient of skewness	1.66	1.66	2.28
Correlation with record	0.91	0.90	1.00

*Streamflow measured in cubic metres per second.

TABLE 2. Comparison between statistics of full rank and reduced rank estimates of flow* at site 02GA018

	Reduced rank estimates	Full rank estimates	Streamflow record
Mean	5.83	5.81	5.55
Standard deviation	7.10	7.16	7.22
Correlation with stn. 1	0.88	0.90	0.89
Correlation with stn. 2	0.82	0.85	0.85
Correlation with stn. 3	0.84	0.84	0.87
Correlation with stn. 4	0.87	0.91	0.93
Correlation with stn. 5	0.86	0.88	0.91
Serial correlation	0.34	0.37	0.34
Coefficient of skewness	1.80	1.91	2.12
Correlation with record	0.89	0.92	1.00

*Streamflow measured in cubic metres per second.

TABLE 3. Comparison between statistics of full rank and reduced rank estimates of flow* at site 02GD001

	Reduced rank estimates	Full rank estimates	Streamflow record
Mean	13.81	13.78	13.59
Standard deviation	13.92	13.92	13.84
Correlation with stn. 1	0.91	0.93	0.97
Correlation with stn. 2	0.75	0.78	0.71
Correlation with stn. 3	0.82	0.83	0.83
Correlation with stn. 4	0.82	0.85	0.80
Correlation with stn. 5	0.83	0.84	0.81
Serial correlation	0.40	0.42	0.43
Coefficient of skewness	1.50	1.55	1.75
Correlation with record	0.91	0.92	1.00

*Streamflow measured in cubic metres per second.

model depends primarily upon the accuracy with which the means, standard deviations, and correlations can be estimated. A cutoff of approximately 95% gives the best results from a reduced rank model for southern Ontario. As there is very little difference between the full rank and reduced rank estimates and as the reduced rank model eliminates the computational problem resulting from highly inter-correlated predictor variables, the use of the reduced rank model is justified.

FULLER, G. A. 1978. Generation of ungauged streamflow data. ASCE Journal of the Hydraulics Division, **104**(HY3), pp. 377–384.

YOUNG, G. K., ORLOB, G. T., and ROESNER, L. A. 1970. Decision criteria for using stochastic hydrology. ASCE Journal of the Hydraulics Division, **96**(HY4), pp. 911–926.

Computer-aided design of building structures—a guide to data center service in Canada

DANIEL J. CARSON AND GILBERT A. HARTLEY

Department of Civil Engineering, Carleton University, Ottawa, Ont., Canada K1S 5B6

Received February 13, 1978

Accepted May 29, 1979

The service available from data centers to structural consultants engaged in the design of buildings is reviewed here. This review includes the main hardware provided by the centers, the application software available, and some comments on the support offered for their application software. This information was obtained by purchasing computer time from data centers in the Ottawa area and using their computer programs in hypothetical design situations. This was supplemented by interviews with data center representatives and a study of brochures and manuals published by the data centers. Although the service available is not of consistently high quality right across the board, there are a few data centers that do offer good service to their engineering clients.

L'article passe d'abord en revue les services fournis par les centres de traitement de données aux ingénieurs en structures oeuvrant dans le domaine du bâtiment. Ce tour d'horizon touche les principaux équipements et les logiciels d'application disponibles dans les centres et se complète de commentaires sur les appuis logistiques correspondants. On a obtenu cette information en achetant des services d'ordinateur de centres de traitement situés dans la région d'Ottawa et en exploitant les programmes d'ordinateur fournis dans des situations de calculs hypothétiques. On a ensuite complété ces sources par des interviews de représentants de centres de traitement et une étude des brochures et manuels publiés par les centres. Quoique les services ne soient pas partout d'une qualité uniformément élevée, il existe quelques centres de traitement des données qui offrent de bons services à leurs clients ingénieurs.

[Traduit par la revue]

Can. J. Civ. Eng., **6**, 472–480 (1979)

0315-1468/79/030472-09$01.00/0